Exemple d'arbre de décision

Bassem Ben Hamed

Juillet 2018

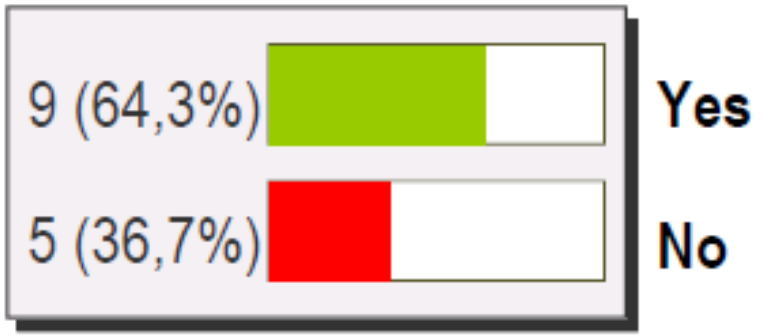
N°	Outlook	Temperature	Humidity	Windy	Play?
1	Sunny	hot	high	false	No
2	Sunny	hot	high	true	No
3	overcast	hot	high	false	Yes
4	rain	mild	high	false	Yes
5	Rain	cool	normal	false	Yes
6	rain	cool	normal	true	No
7	overcast	cool	normal	true	Yes
8	sunny	mild	high	false	No
9	sunny	cool	normal	false	Yes
10	rain	mild	normal	false	Yes
11	sunny	mild	normal	true	Yes
12	overcast	mild	high	true	Yes
13	Overcast	hot	normal	false	Yes
14	rain	mild	high	true	No

Objectif :

- **Prédire si un match de foot va avoir lieu ou non.**
- **Établir une relation** entre le fait de jouer ou pas et les conditions météorologiques.
- **Variable à expliquer (cible) :** Play (2 classes yes et no).
- **Variables explicatives :** Outlook, Temperature, Humidity et Windy

Nœud racine de l'arbre

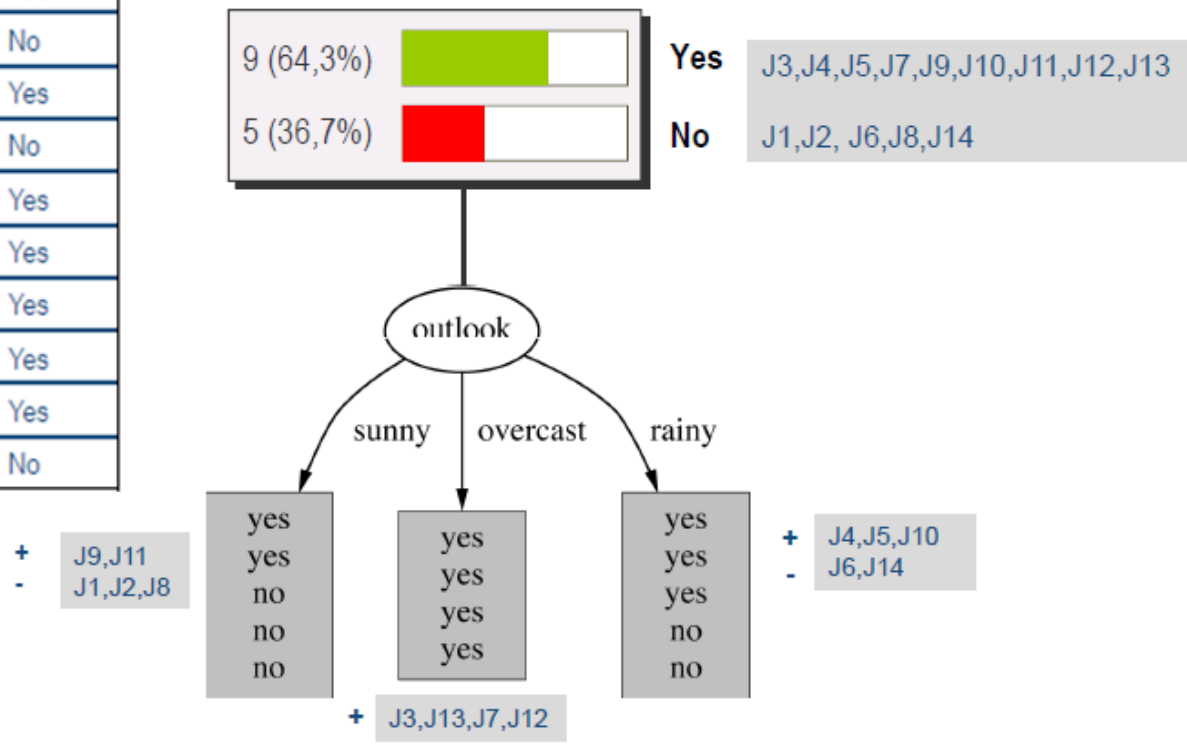
N°	Outlook	Temperature	Humidity	Windy	Play?
1	Sunny	hot	high	false	No
2	Sunny	hot	high	true	No
3	overcast	hot	high	false	Yes
4	rain	mild	high	false	Yes
5	Rain	cool	normal	false	Yes
6	rain	cool	normal	true	No
7	overcast	cool	normal	true	Yes
8	sunny	mild	high	false	No
9	sunny	cool	normal	false	Yes
10	rain	mild	normal	false	Yes
11	sunny	mild	normal	true	Yes
12	overcast	mild	high	true	Yes
13	Overcast	hot	normal	false	Yes
14	rain	mild	high	true	No



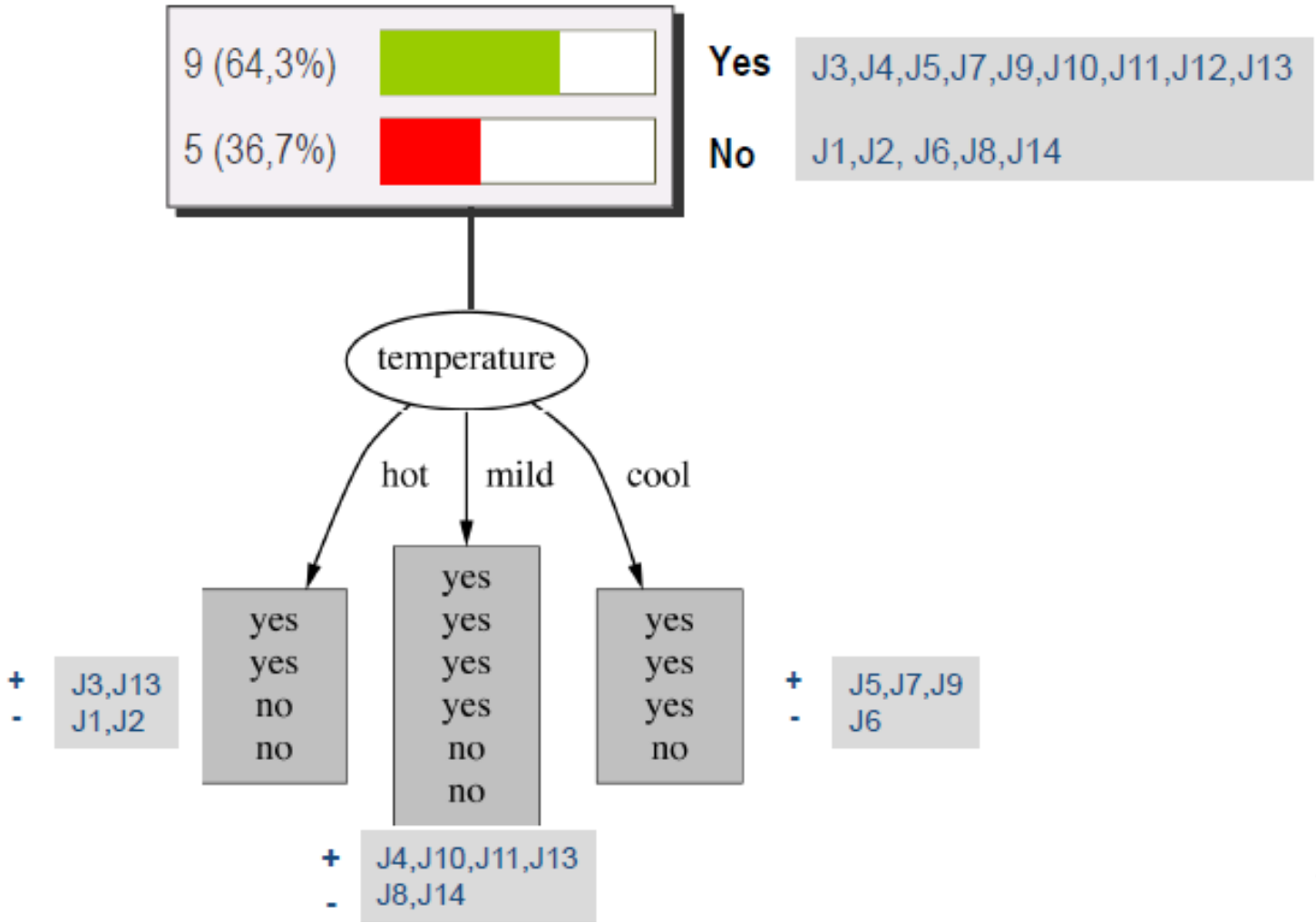
Le nœud racine comprend tous les individus de la base d'apprentissage partitionnés selon la classe à prédire (variable cible).

Comment éclater le nœud racine ?

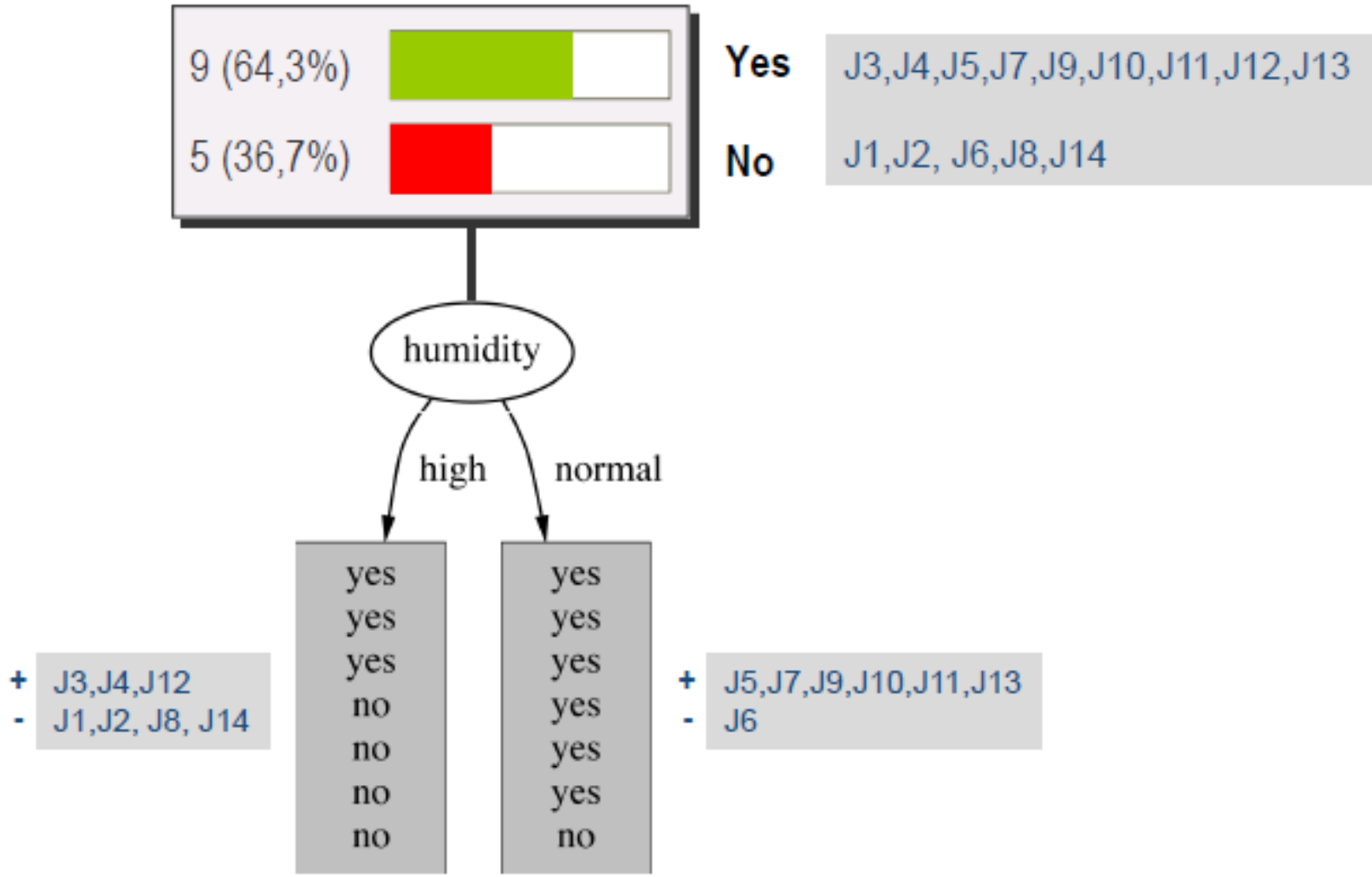
N°	Outlook	Temperature	Humidity	Windy	Play?
1	Sunny	hot	high	false	No
2	Sunny	hot	high	true	No
3	overcast	hot	high	false	Yes
4	rain	mild	high	false	Yes
5	Rain	cool	normal	false	Yes
6	rain	cool	normal	true	No
7	overcast	cool	normal	true	Yes
8	sunny	mild	high	false	No
9	sunny	cool	normal	false	Yes
10	rain	mild	normal	false	Yes
11	sunny	mild	normal	true	Yes
12	overcast	mild	high	true	Yes
13	Overcast	hot	normal	false	Yes
14	rain	mild	high	true	No



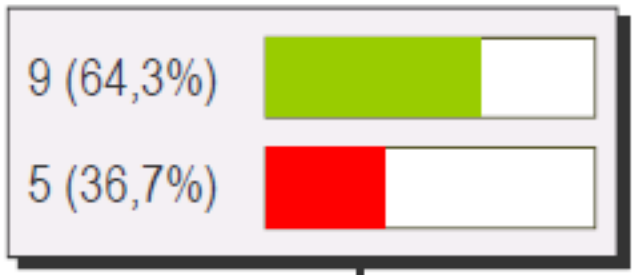
Comment éclater le nœud racine ?



Comment éclater le nœud racine ?



Comment éclater le nœud racine ?



Yes J3,J4,J5,J7,J9,J10,J11,J12,J13

No J1,J2, J6,J8,J14

windy

false

true

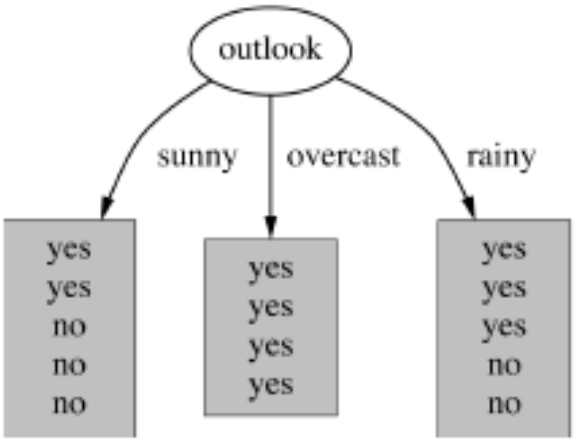
yes
yes
yes
yes
yes
yes
no
no

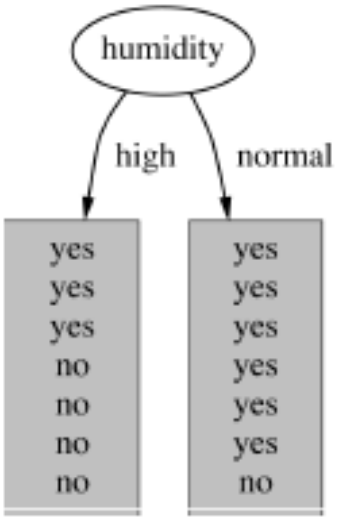
yes
yes
yes
no
no
no

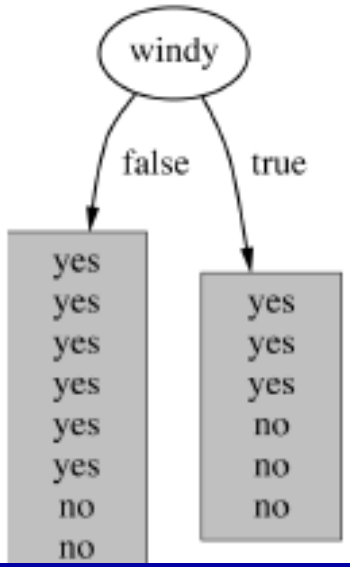
\+ J3,J4,J5,J9,10,J13
\- J1,J8

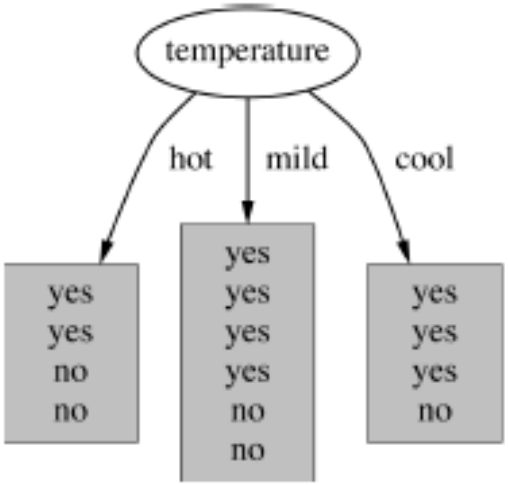
\+ J7,J11,J12
\- J2,J6,J14

Quelle est la variable à choisir ?







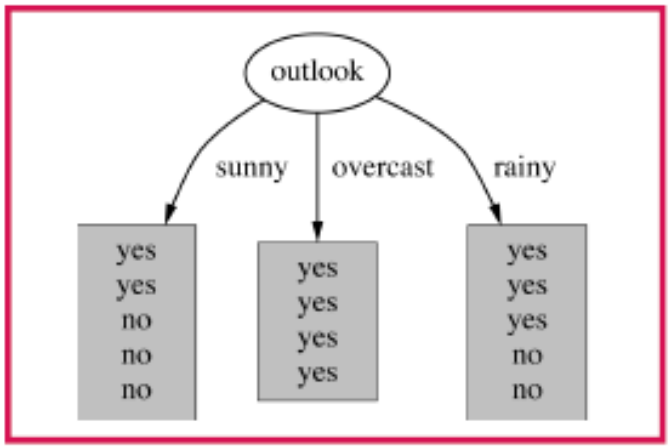


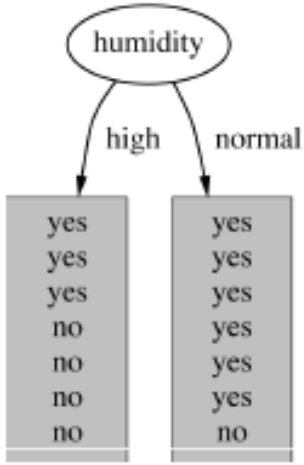
Quelle est la variable à choisir ?

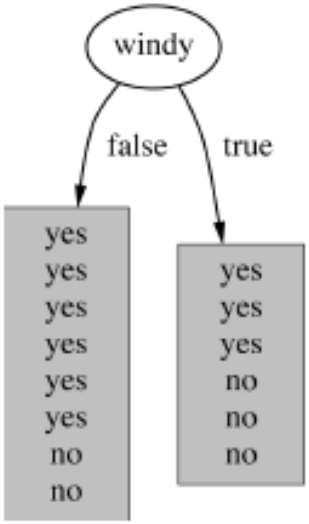
Il faut choisir la variable qui :

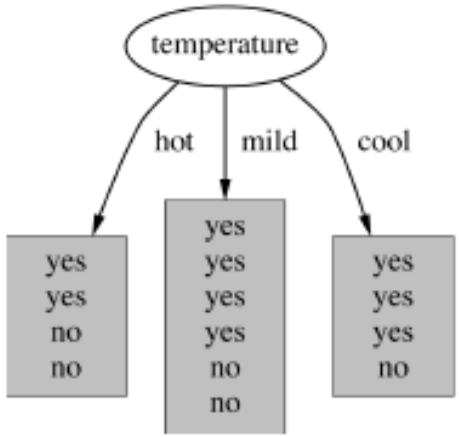
- mène à la création de nœuds fils les plus purs possible.
- diminue le plus possible le désordre (l'entropie) de la classe à prédire dans les nœuds fils.
- mène à une nouvelle partition d'individus qui diminue l'entropie en cours.

Quelle est la variable à choisir ?

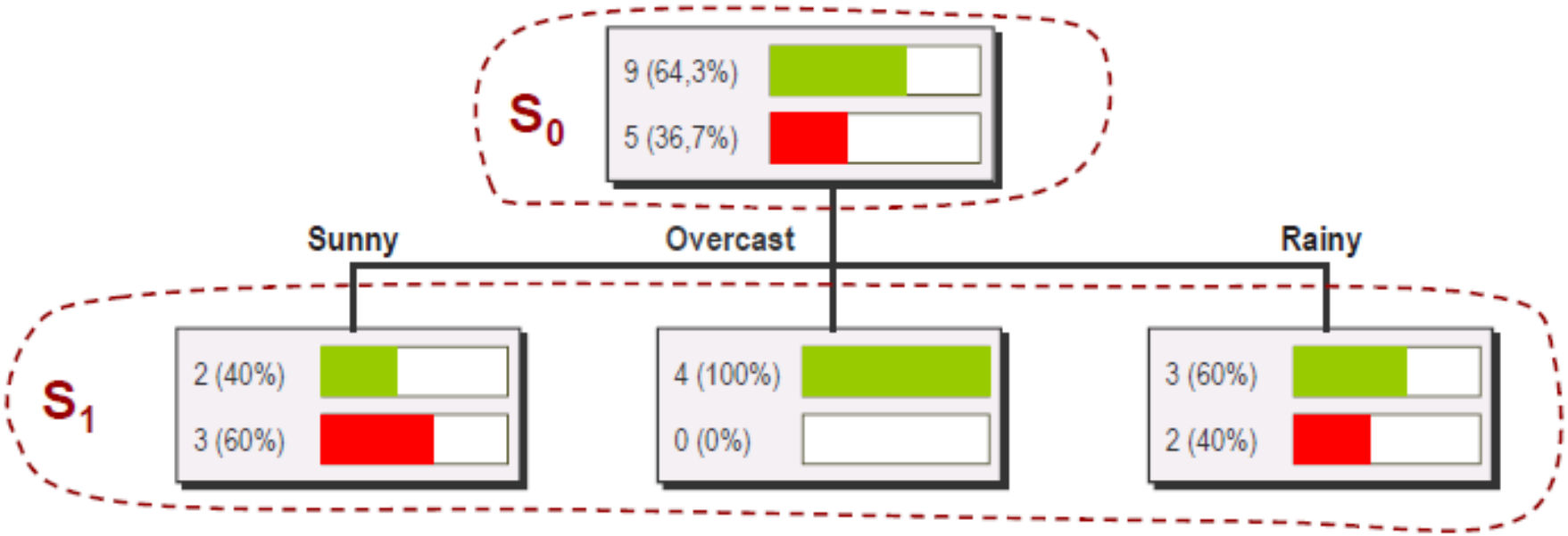




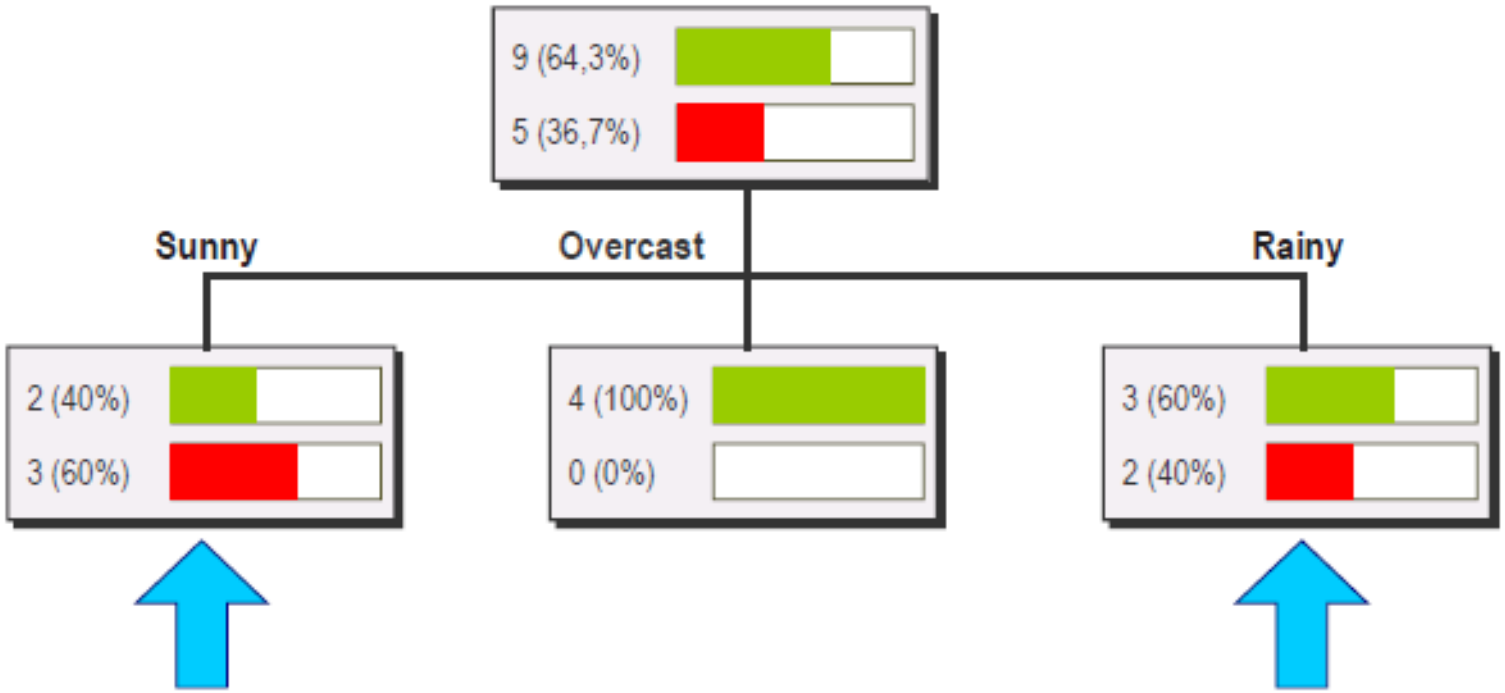




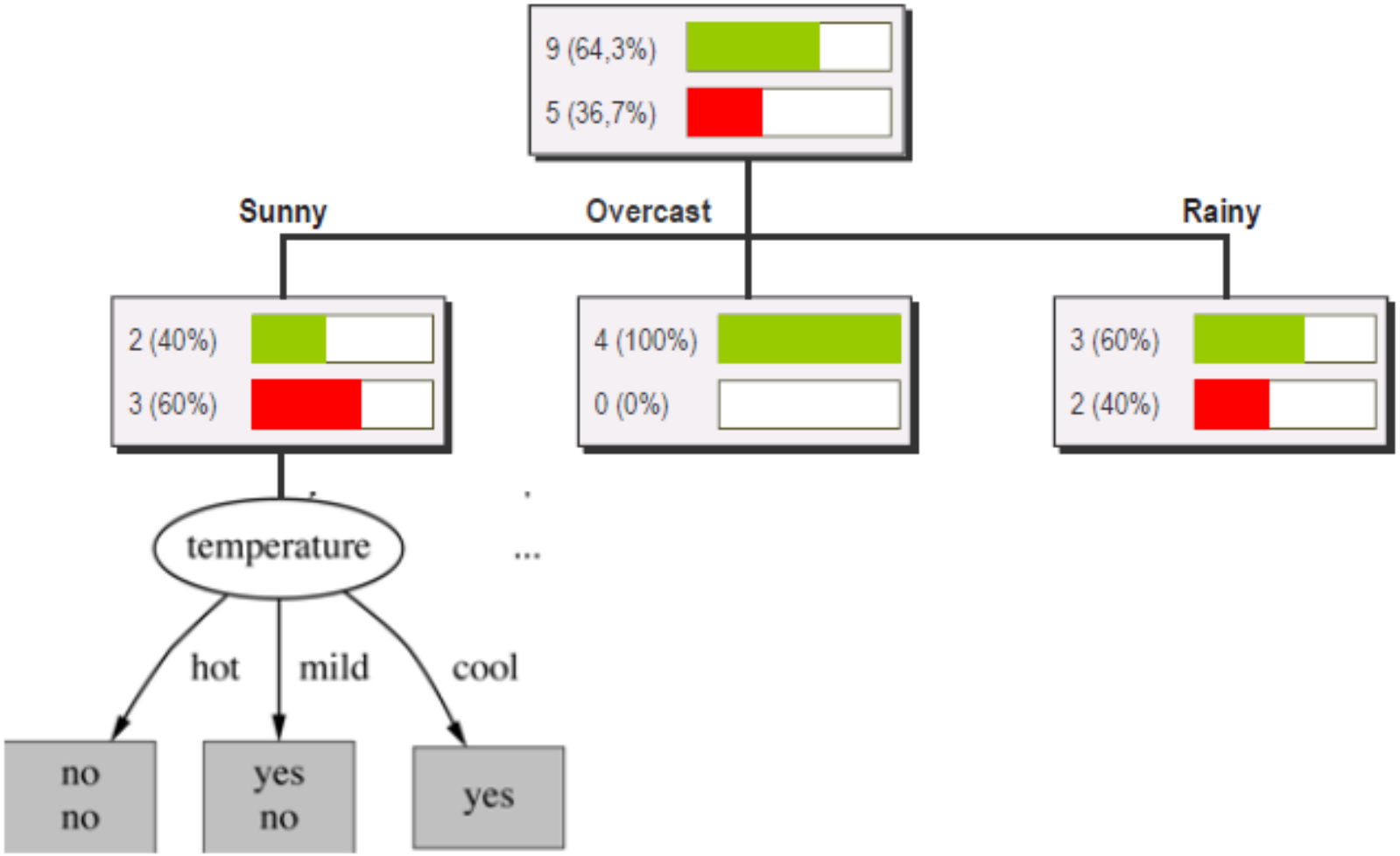
Deuxième partition de l'arbre



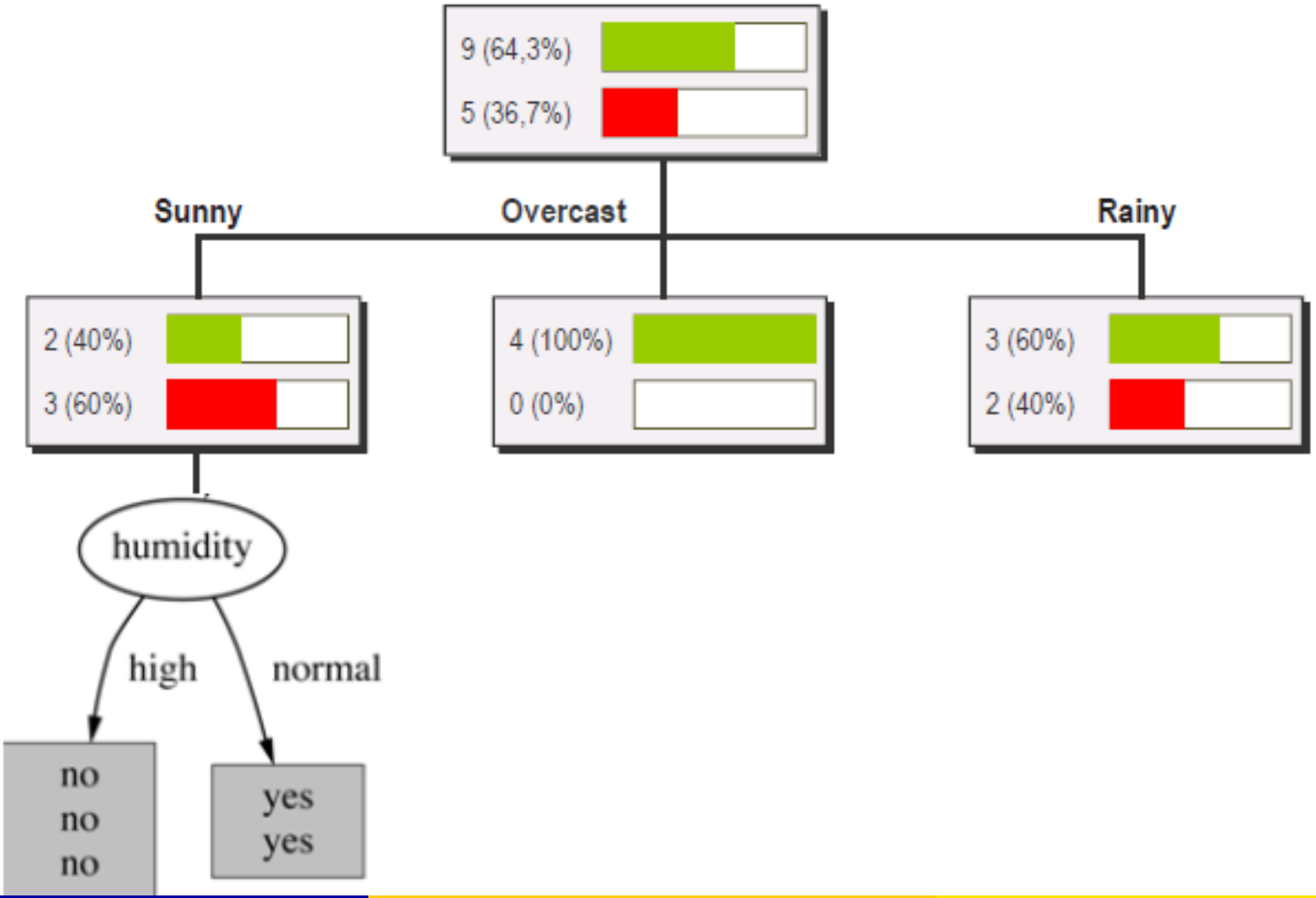
Quel est le nœud à éclater ?



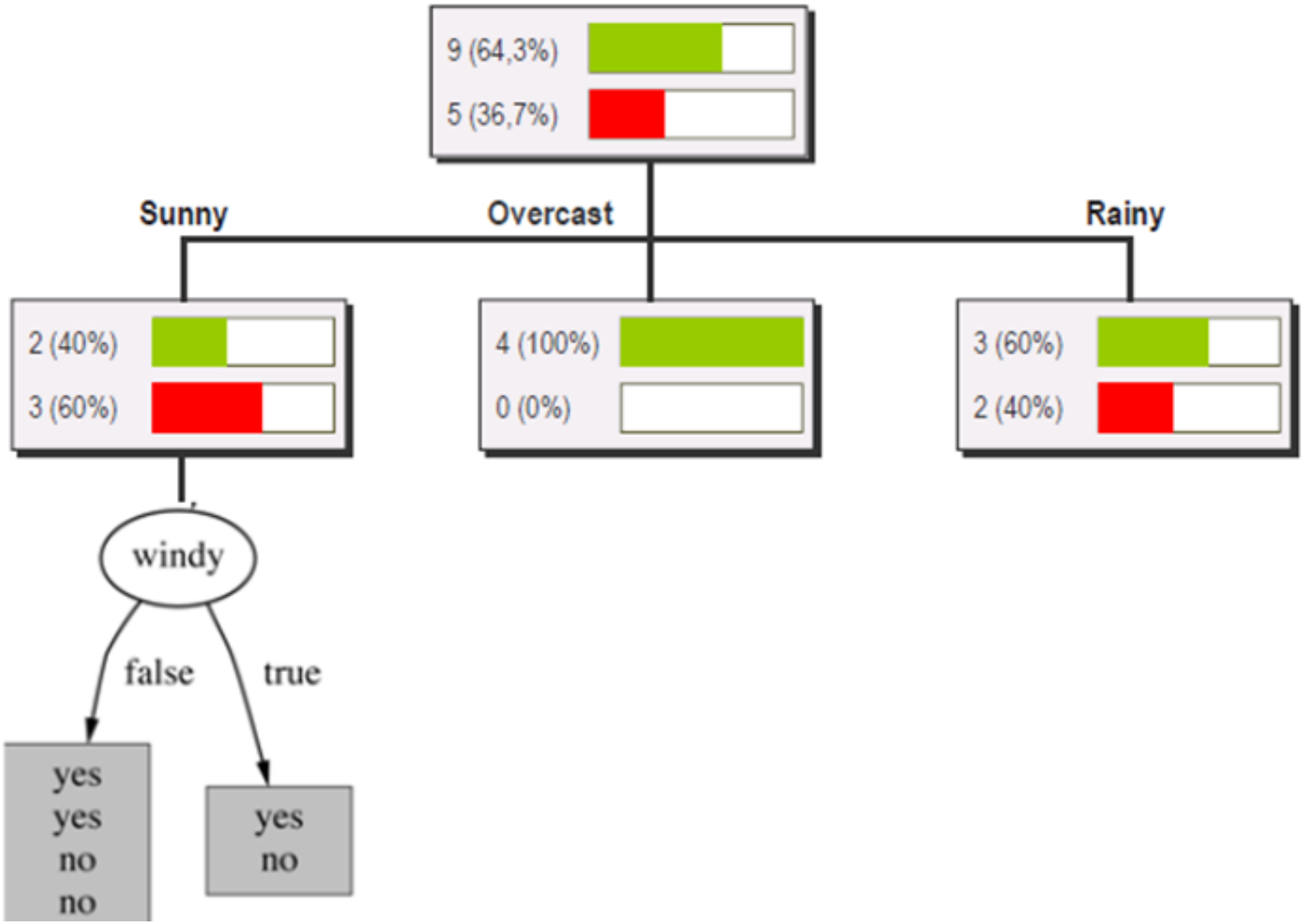
Quelle est la variable à choisir ?



Quelle est la variable à choisir ?

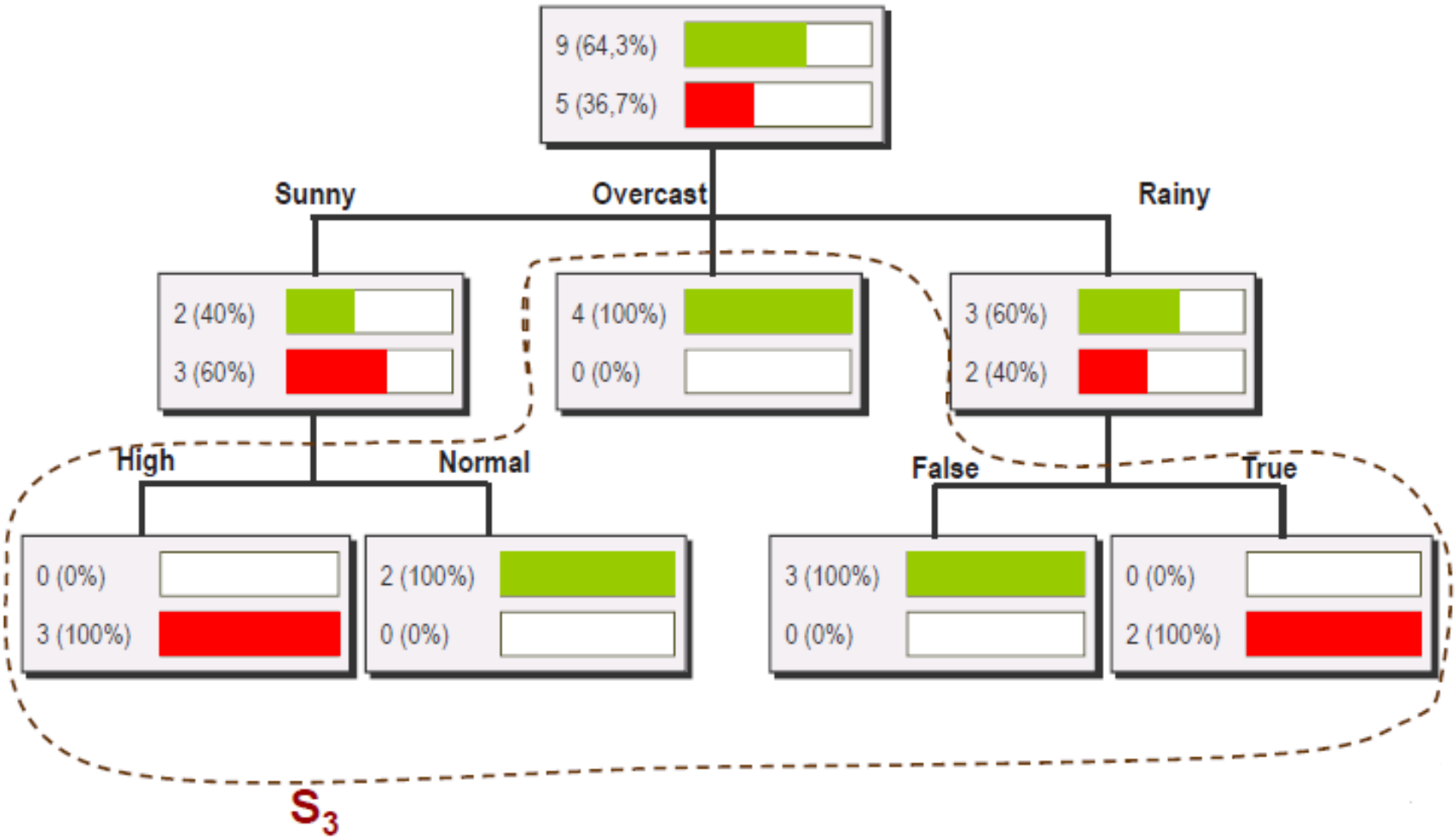


Quelle est la variable à choisir ?



Troisième partition de l'arbre

Quatrième partition de l'arbre



Mesure d'impureté

- Le critères de choix de chaque nœud est la notion de mesure d'impureté

Cette mesure doit :

- être égale à zéro pour un nœud pur de l'arbre de décision
- être croissante en fonction du désordre d'un nœud. Plus le désordre est grand, plus la valeur de la mesure est grande.
- avoir des valeurs additives pour évaluer le désordre d'une partition de l'arbre de décision.

⇨ Entropie de Shannon
⇨ Entropie de Boltzmann
⇨ Index de Gini

Entropie de Shannon

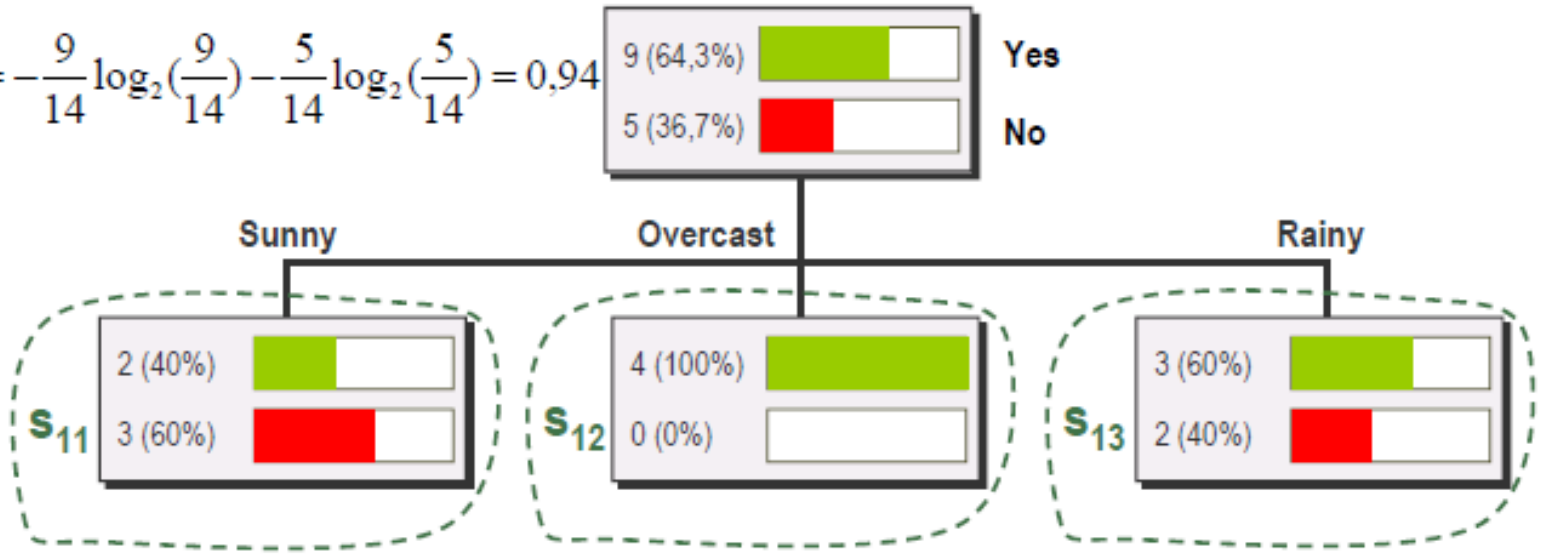
- Shannon en 1949 a proposé une mesure d'entropie valable pour les distributions de probabilité.
- Pour un nœud s, l'entropie d'information est :

$$I(s) = -\sum_{i=1..k} p_i \times \log_2(p_i)$$

où pi est la probabilité de la classe Ci.

Entropie de Shannon

$$I(s_0) = -\frac{9}{14}\log_2(\frac{9}{14}) - \frac{5}{14}\log_2(\frac{5}{14}) = 0{,}94$$



$$I(s_{11}) = -\frac{2}{5}\log_2(\frac{2}{5}) - \frac{3}{5}\log_2(\frac{3}{5}) = 0{,}97$$

$$I(s_{12}) = -\frac{4}{4}\log_2(\frac{4}{4}) - \frac{0}{4}\log_2(\frac{0}{4}) = 0$$

$$I(s_{13}) = -\frac{3}{5}\log_2(\frac{3}{5}) - \frac{2}{5}\log_2(\frac{2}{5}) = 0{,}97$$

NB

$Log_2(x) = Log(x) / Log(2)$

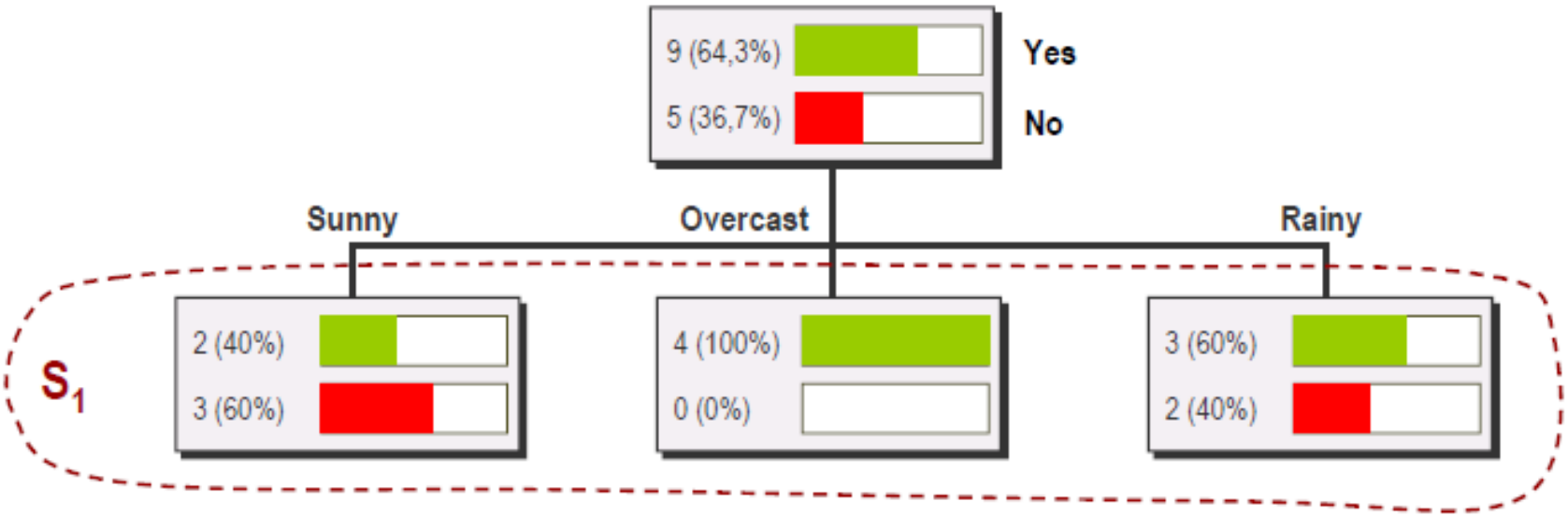
Entropie de Shannon

- Pour une partition S l'entropie d'information est :

$$I(S) = \sum_{s \in S} \frac{Card(s)}{Card(\Omega)} I(s)$$

où I(s) est l'entropie d'information du nœud s

Entropie de Shannon



$$I(S) = \frac{5}{14} I(s_{11}) + \frac{4}{14} I(s_{12}) + \frac{5}{14} I(s_{13})$$

Entropie de Shannon

Critère de partitionnement

- Gain d'incertitude:

$$\Im(S_{t+1}) = I(S_t) - I(S_{t+1})$$

Objectif : Maximiser le gain d'incertitude

- Un nœud p est terminal si : tous les éléments associés à ce nœud sont dans une même classe ou si aucun test n'a pu être sélectionner

Entropie de Shannon

Pour les exemples initiaux
$I(S) = - 9/14 \log2(9/14) - 5/14 \log2(5/14)$

Entropie de l'arbre associé au test sur Outlook ?

- $E(Outlook) = 5/14\ I(S11) + 4/14\ I(S12) + 5/14\ I(S13)$

- Gain(Outlook) = 0.940 - 0.694 = 0.246 bits
- Gain(Temperature) = 0.029 bits
- Gain(Humidity) = 0.151 bits
- Gain(Windy) = 0.048 bits

Choix de l'attribut Outlook pour le premier test

Arbre final obtenu